Royal Signals and Radar Establishment

Memorandum 4152

EXPLICIT MODELLING OF STATE DURATION CORRELATIONS IN HIDDEN MARKOV MODELS

M J Russell and L Sime

September 1988

ABSTRACT

In recent years considerable effort has been directed towards improving the treatment of durational structure in hidden Markov model (HMM) based approaches to speech pattern modelling. In general these studies have been concerned with more accurate modelling of the variations in segment duration which occur when words are spoken at a nominally constant speaking rate. However, recent work has shown that some of the performance gains which can be achieved by improved duration modelling are lost when the words in the test set are spoken at a different rate from those in the training set.

This memorandum presents an approach to solving this problem based on the capture and use of information about state duration correlations. A method for measuring correlations between the durations of adjacent states in a HMM is described. The method involves expanding a standard HMM into a special type of hidden semi-Markov model (HSMM), called a Correlated Duration HMM (CDHMM), in which each state of the original HMM is expanded into a set of fixed-duration HSMM states. The probabilities associated with transitions between these states are measures of state duration correlation.

Experiments are described in which the CDHMM method is applied to a set of sentences spoken at four different speaking rates.

Note: This memorandum is an expanded version of a paper presented at SPEECH '88 – Seventh FASE Symposium, 22–26 August 1988, Edinburgh.

Royal Signals and Radar Establishment

Memorandum 4152

EXPLICIT MODELLING OF STATE DURATION CORRELATIONS IN HIDDEN MARKOV MODELS

M J Russell and L Sime

September 1988

ABSTRACT

In recent years considerable effort has been directed towards improving the treatment of durational structure in hidden Markov model (HMM) based approaches to speech pattern modelling. In general these studies have been concerned with more accurate modelling of the variations in segment duration which occur when words are spoken at a nominally constant speaking rate. However, recent work has shown that some of the performance gains which can be achieved by improved duration modelling are lost when the words in the test set are spoken at a different rate from those in the training set.

This memorandum presents an approach to solving this problem based on the capture and use of information about state duration correlations. A method for measuring correlations between the durations of adjacent states in a HMM is described. The method involves expanding a standard HMM into a special type of hidden semi−Markov model (HSMM), called a Correlated Duration HMM (CDHMM), in which each state of the original HMM is expanded into a set of fixed−duration HSMM states. The probabilities associated with transitions between these states are measures of state duration correlation.

Experiments are described in which the CDHMM method is applied to a set of sentences spoken at four different speaking rates.

Note: This memorandum is an expanded version of a paper presented at SPEECH '88 − Seventh FASE Symposium, 22−26 August 1988, Edinburgh.

1. INTRODUCTION

In recent years considerable effort has been directed towards improving the treatment of durational structure in hidden Markov model (HMM) based approaches to speech pattern modelling [1], [2], [3], [5]. This has resulted in a number of extensions of the standard HMM formalism, such as hidden semi-Markov models (HSMMs) or variable duration HMMs [1], [3], [5], expanded state HMMs [5] and other more pragmatic techniques for incorporating constraints based on observed state duration into the recognition process [4]. In general, these studies have been concerned with more accurate modelling of the variations in segment duration which occur when words are spoken at a nominally constant speaking rate. Under these conditions, results have been presented which confirm that better modelling of durational structure leads to improvements in recognition accuracy [5]. With the exception of [2], there has been little reported work on the effect of speaking rate on the performance of these algorithms.

The results presented in [2] show that some of the performance gains which can be achieved by improved duration modelling are lost when the words to be recognised are spoken at a different rate from those in the training set. The solution proposed in [2] is based on the hypothesis that although absolute durational structure will be affected, relative durational structure is invariant to changes in speaking rate. Using the measures of likelihood of state duration ratio developed in [2] some success in recognising speech spoken at different rates was achieved, but the problem was not completely overcome.

One possible solution is to build a HSMM using examples of speech spoken at a representative range of speaking rates. However this will result in models in which the state duration probability density functions (pdfs) have large variances, and such a model will not in general be able to discriminate between words, like 'pod' and 'pot', which are spectrally similar and rely on the detection of significant temporal contrasts for recognition.

The motivation for the method presented in this memo is the assumption that what is missing in the simple approach just described is the ability to capture and make use of information about state duration correlations. In order to decide whether, for example, a short state duration is acceptable because the speaker is speaking quickly, or unacceptable because he is not, it is necessary to know the probability of this state duration conditioned on the durations spent in previous states.

The purpose of the present memorandum is to describe and evaluate a method for measuring correlations between the durations of successive states directly. The method involves expanding a standard HMM into a special type of HSMM, which is called a *Correlated Duration* HMM (CDHMM), such that each state in the original HMM is expanded into a set of fixed-duration HSMM states. The probabilities associated with transitions between these states are measures of state duration correlation.

The technique has been applied to a phonetically balanced set of sentences recorded at four different speaking rates, 'normal', 'fast', 'very fast' and 'fast with rate meter' [6], by two female and two male subjects. Four applications of the results are then considered. First the results are used to examine how the strategies which the speakers use to change their speaking rates are reflected in terms of changes in state duration. Second, the potential usefulness of constraints based on local state duration correlations for rate-independent speech recognition is investigated. The third application is a search for any abnormal durational effects which are introduced by the use of the 'rate meter' to control speaking rate. Finally the results are used to test the validity of the linear model of state duration correlation proposed in [2]. The main conclusion is that although the results demonstrate that the CDHMM method is able to capture local state duration correlations, they suggest that these local correlations may not be sufficient to solve the problem of recognising speech spoken at different rates and that what is needed is the ability to model correlations in the durations of (not necessarlily adjacent) phonetically related regions.

The memorandum is arranged as follows. HMM and HSMM terminology and notation is established in section 2. The method for deriving a CDHMM from a standard HMM is described in section 3, and issues of model parameter estimation for CDHMMs are addressed briefly in section 4. Section 5 describes the experiments in which the technique was used to analyse sentences spoken at different speaking rates, and the results of the experiments are presented and discussed in sections 6, 7 and 8. Some conclusions are presented in section 9.

2. DURATION MODELLING IN HIDDEN MARKOV MODELS

2.1 Hidden Markov Models

The assumption behind HMM based approaches to speech pattern processing is that a sequence of acoustic observations, $Y_1,\ldots,Y_T$, which represents a given utterance can be modelled as a probabilistic *function* of a *finite state* Markov *chain* $X_1,\ldots,X_T$. The structure of this Markov chain and its relationship with the sequence of acoustic observations is defined by an N-state HMM $M = (\pi,A,b)$, where, if the i^{th} state of M is denoted by s_i,

(2a) $\pi_i = Prob(X_1{=}s_i)\quad i{=}1,\ldots,N$. The vector $\pi = (\pi_1,\ldots,\pi_N)$ is called the *initial state probability vector.*

(2b) $A = [a_{ij}]_{i,j=1,..,N}$ is a row-stochastic matrix such that $a_{ij} = Prob(X_t{=}s_j \mid X_{t-1}{=}s_i)$. A is called the *state transition probability matrix.*

(2c) $b = b_1,\ldots,b_N$, and b_i is a pdf defined on the set of observations such that, for any Z, $b_i(Z) = Prob(Y_t = Z \mid X_t = s_i)$ $i{=}1,\ldots,N$. The pdf b_i is called the *state output pdf* associated with state s_i.

Such a model assumes that a speech pattern can be represented as the output of a sequence of stationary stochastic processes of varying duration. The individual processes are defined by the pdfs associated with the states of the HMM, and the underlying Markov process models durational and sequential structure.

Although the assumption of piecewise stationarity is clearly an approximation in the context of speech pattern modelling, the limitations of HMMs are offset to a significant degree by the existence of a computationally useful mathematical theory of HMMs which includes powerful algorithms for automatic HMM parameter estimation from data, and for HMM-based speech pattern recognition [10], [11]. In particular, given a set of training patterns $O^1,\ldots,O^S$ and an initial estimate of a suitable set of HMM parameters, Baum's theorem provides the basis for an iterative algorithm which locally maximises the likelihood $Prob(O^1,\ldots,O^S \mid M)$ [10].

2.2 Improved duration modelling using hidden semi-Markov models

It follows from the definition in section 2.1 that state duration in a HMM is modelled as an exponentially decaying geometric pdf which assigns maximum probability to a duration of one time unit and progressively smaller probabilities to longer durations [7]. The unsuitability of this class of pdfs as a model of segment duration in speech patterns has prompted the consideration of HMM-type models where the underlying Markov process is replaced by a *semi*-Markov process in which state duration is modelled explicitely. The resulting model is called a hidden semi-Markov model (HSMM) [1], [5], [7], or Variable Duration HMM [3]. Formally, a HSMM is simply a HMM in which each state s_i is associated with a *state duration pdf* d_i such that $d_i(D)$ is the probability that s_i is occupied for exactly D time units. Automatic HSMM parameter reestimation from training data, using extensions of the Viterbi and Baum-Welch reestimation procedures, is valid for a range of classes of state duration pdf [1], [3], [8].

In the case where the state duration pdfs are non-parametric, the resulting HSMM is referred to as the Ferguson Model [6]. This type of HSMM is particularly relevant for the present study.

3. CORRELATED DURATION HMMS

3.1 Correlated duration hidden Markov models

Let $M = (\pi,A,b)$ be an N state HMM, and let D_{max} be the maximum allowable state duration. The *Correlated Duration HMM* (CDHMM) associated with M is an $N.D_{max}$ state HSMM $M^* = (\pi^*,A^*,b^*,d^*)$. Indexing the states of M^* by pairs (i,x) $(i=1,\ldots,N;$ $x=1,\ldots,D_{max})$, where i corresponds to a state in the original HMM M and x denotes duration in that state, M^* is specified as follows:

(3a) $b^*_{(i,x)} = b_i$ for all values of x.

(3b) $a^*_{(j,x)(j,y)} = a_{ij}/(D_{max}(1-a_{ii})),\ i{\neq}j.$ In particular, if $a_{ij} = 0$, then $a^*_{(i,x)(j,y)} = 0$ for all values of x and y.

(3c) $a^*_{(i,x)(i,y)} = 0$ for all x,y.

(3d) $\pi^*_{(i,x)} = \pi_i/D_{max}$. In particular, if $\pi_i = 0$, then $\pi^*_{(i,x)} = 0$ for all x.

(3e) $d^*_{(i,x)}(y) = 1$ if $x = y$, 0 otherwise.

The relationship between a simple left-right HMM M and its associated CDHMM M^* is shown in figure 1. Each state s_i of M is expanded into a column of D_{max} states which have the same state output pdf $b^*_{(i,x)} = b_i$, but different state duration pdfs: the x^{th} state in the column can only be occupied for exactly x time units because of the definition of $d^*_{(i,x)}$. Hence, if the parameters of M^* are reestimated from training data (see section 4), $\pi^*_{(i,x)}$ will be the probability that s_i is the initial state and is occupied for exactly x time units, and $a^*_{(i,x)(j,y)}$ will be the probability of a duration of y time units in state s_j, given that a duration of x time units in state s_i occured immediately before. In other words it is assumed that the probability of a particular state duration depends only on the duration of the immediately preceding state.

Notice that if the model is constrained so that $a^*_{(i,x)(j,y)} = a^*_{(k,z)(j,y)}$, for all i, j, k, x, y and z, then M^* reduces to a normal Ferguson type HSMM.

3.2 Relationship between HMM state sequences and paths through the CDHMM network

Figure 2 shows the correspondence between a state sequence $X = X_1,\ldots,X_T$, which relates an observed sequence $Y = Y_1,\ldots,Y_T$ to the HMM M, and a path through the network corresponding to M^*. Since this type of figure occurs several times in the memorandum, it is useful to describe it in detail. The spectrogram at the bottom of the figure represents an example of an utterance of the digit 'zero', it is one of the patterns that was used to train the HMM M. The spectrogram immediately above is a 'synthesised' version of 'zero' which consists of the mean vectors of the states in the sequence X: the t^{th} vector in the pattern is the expected output of state X_t. In this and subsequent figures the state sequence X is the sequence which maximises the probability $Prob(X,Y|M)$, hence the 'synthetic' spectrogram can be interpreted as the best explanation of the 'real' spectrogram in terms of the HMM. The 'cones' in the figure simply show how the states of the HMM map onto the synthesised spectrogram.

This type of picture is extremely useful for interpreting the states of a HMM in terms of acoustic events, however it should be noted that it does not include information about the variance of the state output pdfs.

4. PARAMETER ESTIMATION FOR CDHMMS

4.1 Baum–Welch parameter reestimation for CDHMMs

A CDHMM is a Ferguson type HSMM [5] in which constraints are placed on the forms of the state duration pdfs. It follows that standard HSMM parameter reestimation techniques, modified to ensure that condition (3a) is not violated, are applicable to CDHMMs [8]. This approach to parameter estimation was adopted in the experiments.

4.2 Undertraining in CDHMMs

Because the transformation of an HMM into a CDHMM involves a factor of $D_{max}{}^2$ increase in the number of state transitions, there is a real danger that the model will be undertrained if the straightforward approach to model parameter estimation described above is followed. This could be alleviated to some extent by parameterising the pdf $a^*_{(i,x)(j,-)}$, using for example binomial pdfs [2], or by some other form of smoothing.

5. EXPERIMENTS WITH CDHMMS USING SENTENCES SPOKEN AT DIFFERENT SPEAKING RATES

5.1 Choice of speech data

The CDHMM method was investigated using recordings of a phonetically balanced set of sentences spoken by four speakers (2 female and 2 male). The sentences are those specified in the SPAR database and are listed in appendix 1. Two of the speakers (RM (male) and SJ (female)) are members of staff at the Speech Research Unit (SRU), the remaining speakers (HP (female) and CR (male)) were vacation students at SRU. Each speaker spoke 30 examples of each sentence at 'normal' speaking rate, and 20 examples of each sentence at 'fast', 'very fast' and 'fast with rate meter' speaking rates. Sentences were chosen in preference to individual words because it was felt that it would be unrealistic for subjects to pronounce isolated words naturally at different speaking rates.

5.2 Recording procedure

Randomly ordered sentence lists were presented to the speakers on a VDU display, which also showed a level meter and, in the final condition, a rate meter. The later was a moving cursor, with brightness proportional to speech level, which described a circle on the VDU screen. The prompting and recording system and the rate meter are described in detail in [6].

5.3 Definition of the different speaking rates

The four different speaking rates were defined as follows: Under the 'normal' condition speakers were instructed to speak the sentences at what they considered to be their normal rate. For the 'very fast' condition speakers were asked to speak as fast as they could whilst still remaining intelligible. The 'fast' condition was an intermediate between 'normal' and 'very fast'. Under the 'fast with meter' condition, speakers were asked to speak sufficiently quickly so that, on average, the visible cursor described a semi–circle on the VDU screen. The speed of the cursor was set so that the time taken to complete a semi–circle was two–thirds of the average duration of the sentences spoken normally by the speaker. All four of the conditions were explained to the subjects and practiced by the subjects before any recordings were made. The terms 'normal', 'fast', 'very fast' and 'fast with meter' will be used in this context throughout the remainder of the memorandum.

5.4 Pre–processing

The recordings were processed using the SRU standard SRUbank filterbank analyser software configured to its default setting of 27 filters, spanning the range up to 10KHz, but the frame rate was reduced from the default value of 200 frames per second to 100 frames per second. The processed recordings were annotated orthographically at the sentence level using a semi–automatic system with manual verification.

5.5 Sentence level hidden Markov models

Speaker-dependent HMMs were constructed for each sentence spoken at normal speaking rate: All examples of a given sentence spoken at normal rate by a given speaker were time-aligned, using a standard dynamic time-warping algorithm, and combined into a single composite pattern. This was segmented using the dynamic programming algorithm described in [9], and the segments were used to estimate the parameters of a HMM. The parameters of the resulting model were reestimated using the Baum-Welch algorithm, applied to the the same set of training sentences, to produce an optimised HMM M.

In all experiments the HMM state output pdfs b_i were single multivariate Gaussian with diagonal covariance matrices. The underlying Markov model was a simple left-right model in which only transitions from state s_i to states s_i and s_{i+1} were assigned non-zero pobability. This particularly restrictive model topology was chosen so as to minimise the number of parameters in the associated CDHMM. The number of states was varied between 20 and 30.

The lower halves of each of the pictures in figure 3 can be used to interpret the 20 state HMMs of the sentence 'Why are you early, you owl?', spoken by each of the speakers. As in figure 2, the bottom spectrogram represents one of the utterances which were used to train the HMM, the top spectrogram represents the best explanation of this utterance in terms of the parameters of the HMM, and the cones show how the HMM states map onto the spectrogram (see section 3.2). Using this type of figure it is possible to try to interpret the states of the HMM in broad phonemic terms. In figure 3(a), for example, states 3, 6, 9, 13 and 16 are associated with relatively long stationary regions in the speech pattern, corresponding approximately to the initial part of the diphthong /aI/ in 'why', /j/ in 'you', /ɜ/ in 'early', /j/ in 'you' and the initial part of the diphthong /aU/ in 'owl' respectively. The remaining states are used to obtain piecewise stationary approximations to the non-stationary portions of the pattern.

Subjectively the worst model is the one shown in figure 3(d), corresponding to speaker CR. In this model there are clearly several states (for example states 8, 10 and 12) which would have been ommitted if the underlying model topology had been rich enough to allow this. The speech of subject CR was the most variable in the study.

Figure 4 includes the corresponding information for the 20 state HMMs of the sentence 'Six plus three equals nine' for each of the 4 subjects.

Figures 3 and 4 suggest that 20 states are adequate to model the sentences 'Why are you early, you owl?' and 'Six plus three equals nine'. However, 20 states were insufficient for the remaining three sentences. Although 30 state HMMs of these sentences were constructed, the amount of computing time required to process the resulting CDHMMs was excessive. For this reason it was decided to restrict the majority of the experiments to 20 state models and the sentences 'Why are you early, you owl?' and 'Six plus three equals nine'.

5.6 Correlated Duration HMMs at the sentence level

For each speaker and each sentence the HMM M was expanded into a CDHMM M^* using the procedure described in section 3.1. Baum-Welch HSMM reestimation was then used to reestimate the state transition probabilities in M^* [8]. Reestimation was conducted separately for the sets of sentences spoken at normal, fast, very fast and fast with rate meter rates, resulting in 4 CDHMMs per sentence per speaker.

Notice that there is no technical reason why condition 3(a), which specifies that all states in a single vertical column of a CDHMM are associated with the same state output pdf, cannot be relaxed during CDHMM parameter reestimation. The standard HSMM reestimation procedure includes expressions for reestimating the mean and covariance matrix of a multivariate Gaussian state output pdf [8]. This would enable the 'target' spectral parameters associated with a given CDHMM state to adapt to a set of parameters which is appropriate for its associated duration. However, in the experiments reported below condition 3(a) was not relaxed and the state output pdfs were not reestimated during CDHMM optimisation.

6. EFFECTS OF CHANGES IN SPEAKING RATE ON STATE DURATION

6.1 Comparison of 'normal' and 'very fast' rate speech

Figures 3 and 4 show 20 state CDHMMs, with maximum duration D_{max} equal to 15, of the sentences 'Why are you early, you owl?' and 'Six plus three equals nine' spoken by each of the 4 speakers. The red and green levels which are used to draw the link between states $s_{(i,x)}$ and $s_{(j,y)}$ indicate the probability $P_{(i,x)(j,y)}$ of a transition between these two states in normal (red) and very fast (green) rate speech. It follows that transitions which are highly probable for both normal and very fast rate speech appear in yellow. $P_{(i,x)(j,y)}$ is the sum of the probabilities, conditioned on A^*, of all state sequences which include a transition between $s_{(i,x)}$ and $s_{(j,y)}$. This sum is calculated using a forward−backward computation on the semi−Markov model network defined by A^*. More precisely,

$$P_{(i,x)(j,y)} = \gamma(i,x)a^*_{(i,x)(j,y)}\varphi(j,y)$$

where γ and φ are defined by

$$\gamma(i,x) = \sum_{d=1}^{D_{max}} \gamma(i-1,d)a^*_{(i-1,d)(i,x)}$$

and

$$\varphi(i,x) = \sum_{d=1}^{D_{max}} \varphi(i+1,d)a^*_{(i,x)(i+1,d)}$$

subject to the initial conditions $\gamma(1,x) = \pi^*_{(1,x)}$ and $\varphi(N,x) = 1$.

Recall that in each picture the bottom spectrogram Y represents one of the training utterancess, and the top spectrogram is the sequence of mean vectors corresponding to the state sequence X which maximises $Prob(X,Y|M)$, as in figure 2.

Ignoring for the moment figures 3(d) and 4(d), the figures show highly variable durational structure for states which correspond to regions in the speech pattern which are truly relatively stationary (for example states 3, 6, 8, 9 and 13 in figure 3(a)), but much less durational variability for states which map onto non−stationary regions (for example states 5, 7 and 12 in figure 3(a)).

In general, for a fixed speaking rate there is a strong correlation between occurances of the expected durations of adjacent states, but the figure also shows pairs of adjacent 'alternating states', where a short (respectively long) duration in the first state is strongly correlated w.:h a long (respectively short) duration in the second state. An example of a pair of alternating states is states 16 and 17 in figure 3(c) for normal rate speech. These result from an over−constrained A matrix which disallows parallel states. In a standard HSMM this effect would only be observed indirectly through the appearance of multimodal state duration pdfs.

In figures 3(a), (b) and (c) the transitions associated with normal rate speech (red) are generally higher−up than those associated with very fast speech (green), demonstrating that, as one would expect, state durations in very fast speech are shorter than the corresponding state durations in normal rate speech. It is also clear that the differences in state duration statistics are not distributed uniformly over all of the states, but are focussed onto specific states. In the case of figure 3(a) these are states 3, 6, 9 and 13 of the underlying Markov model. It has already been noted that these states correspond to the relatively stationary regions of the speech pattern. This suggests that the primary mechanism used by the speaker to change speaking rate is to shorten or lengthen these regions.

Figures 3 and 4 suggest that correlations between the durations of states which correspond to relatively long stationary regions of the speech pattern, such as states 3, 6, 8, 9 and 13 in 3(a), are the main cues for determining speaking rate. The local state duration correlations captured by the CDHMMs in the present study appear to play a less vital role. For example, again in figure 3(a), a duration of 4 centiseconds in state 12 is strongly correlated with a relatively long duration in state 13 in normal speech, but the same duration correlates with a short duration in state 13 for very fast speech. The problem is that the durations of states like state 12, which correspond to non−stationary regions of the speech pattern, are relatively rate−invariant. This suggests that constraints based on local correlations between the durations of states in this type of HMM may not be sufficient to achieve good speech recognition performance across speaking rates. Running the necessary speech recognition experiments to test this hypothesis is beyond the scope of this memorandum.

Although the above description has concentrated on figure 3(a), similar results can be seen in figures 3(b), 3(c), 4(a), 4(b) and 4(c). However, the normal speech of CR, the 4th speaker in the study is subjectively very fast, and it is not possible from figure 3(d) to find consistent durational differences between his normal and very fast speech. Figure 5 shows graphs of expected state duration as a function of state number for each of the subjects speaking the sentence 'Why are you early, you owl?' at normal and fast speaking rates. In particular figure 5(d) shows that the expected durations of 9 of the 20 states in the model for speaker CR are longer for fast speech than for normal speech.

For completeness, figures 6 to 10 show graphs of expected state duration plotted against state number for all 4 subjects speaking the sentences 'Six plus three equals nine' at normal and fast rates (figure 6), 'Why are you early, you owl?' at normal and very fast rates (figure 7), 'Six plus three equals nine' at normal and very fast rates (figure 8), 'Why are you early, you owl?' at fast and very fast rates (figure 9) and 'Six plus three equals nine' at fast and very fast rates (figure 10).

7. EFFECT OF THE RATE METER

It was thought that the rate meter might cause speakers to adopt unnatural strategies for changing speaking rates. The results of the CDHMM experiments suggest that this is not the case. Figures 11 and 12 show graphs of expected state duration as a function of state number for all 4 subjects speaking the sentences 'Why are you early, you owl?' (figure 11) and 'Six plus three equals nine' (figure 12) at very fast and fast with meter speaking rates. In the case of the first three speakers, the figures show little difference between the patterns of expected state durations for speech spoken very fast and fast with rate meter. The agreement between these two sets of curves was unexpected, since the definitions of the very fast condition, where speakers were asked to speak as quickly as possible while still remaining intelligible, and the fast with meter condition, where the meter was set to obtain an average sentence duration of two−thirds that in normal speech, appear to be very different.

There are several possible explanations. The criteria used to choose the figure 'two−thirds' were basically the same as those used to define the very fast condition. Also, although speakers make considerable use of the rate meter initially to judge an appropriate rate, they are soon able to achieve this rate comfortably and then use the meter only for confirmation. In either case, it appears that the rate meter does not introduce any unnatural effects which are detectable in terms of state duration structure.

8. EFFECT OF CHANGES IN SPEAKING RATE ON STATE DURATION RATIOS

In [2] Kuhn and Ojamaa investigate the extent to which measures of state duration ratio are invariant to changes in speaking rate. One of the measures which they consider is the ratio $R_1(i) = d_i / (d_i + d_{i+1})$, where d_i denotes the duration of the ith state of the HMM. The expected value $E_1(i)$ of $R_1(i)$ is given by the equation

$$E_1(i) = \sum_{1 \leqslant x,y \leqslant D_{max}} Prob(d_{i+1}=y, d_i=x) \frac{x}{x+y} = \sum_{1 \leqslant x,y \leqslant D_{max}} Prob(d_{i+1}=y | d_i=x) . Prob(d_i=x) \frac{x}{x+y}.$$

In a CDHMM, $Prob(d_{i+1}=y | d_i=x)$ is just $a^*_{(i,x)(i+1,y)}$, and $Prob(d_i=x)$ can be calculated using the forward−backward computation described in section 6.1. Hence it is possible to evaluate $E_1(i)$ using the information which is available in a trained CDHMM. Figures 13 and 14 show $E_1(i)$ as a function of i for each of the speakers and the sentences 'Why are you early, you owl? (figure 13) and 'Six plus three equals nine' (figure 14), spoken at both normal and very fast rates. There are clearly substantial differences between the two graphs, suggesting that the ratio E_1 is not rate invariant. A large difference occurs, for example, at state 7 (figure 13) where a state with relatively rate−invariant duration is followed by a state whose duration is highly sensitive to rate (see figure 3 for an interpretation of this state in terms of the speech pattern).

It is likely that the sensitivity to speaking rate of the state duration ratio E_1 which is apparent in figures 14 and 15 is a consequence of the type of model used in the experiments. In the HMMs considered in [2], a state consists of three segments: an initial segment of fixed duration, a centre segment of variable duration, and a final state of fixed duration. These states correspond to relatively long, phonetically identifiable segments of the acoustic pattern consisting of an initial transition, a stationary region and a final transition. Relative invariance of measures like E_1 for these types of states is supported by the correlation between the durations of the relatively long stationary regions observed in section 6.1

9. CONCLUSIONS

This paper has presented CDHMMs as a framework for modelling local state duration correlations in HMMs. Experiments were conducted in which CDHMMs were constructed using examples of sentences spoken at different rates. The resulting models were then used to investigate, in terms of HMM state duration, the strategies which speakers adopt to change their speaking rate, the potential usefulness of measures of local state duration correlation for rate−independent speech recognition, the effect of using the rate meter to control speaking rate, and the invariance of a measure of state duration ratio to changes in speaking rate.

Although the CDHMM method has provided insights into the effects of speaking rate on HMM state duration structure, the results suggest that the extent to which information about local state duration correlations can be useful in achieving good recognition accuracy across speaking rates will depend critically on the nature of the states themselves. In the experiments reported in this memorandum the characteristics of the states were determined by the mathematical techniques which were used to construct the models. Consequently, although some states correspond to phonetically identifiable segments of the acoustic pattern, such as relatively stationary regions in vowel sounds, others are used to obtain piecewise stationary aproximations to non−stationary regions of the speech pattern. These second types of state tend to have relatively rate−invariant durations and hence they are unable to lose important durational cues to speaking rate which are present in previous states. By contrast, the states which form the models considered in [2] are 'composite' states which correspond to relatively large, phonetically significant regions of the speech pattern. The results presented in section 6 suggest that local correlations between the durations of these types of states would provide strong indications of speaking rate.

10. REFERENCES

[1] H BOURLARD and C WELLEKENS, 'Connected digit recognition by phonemic semi-Markov chains for state occupancy modelling', EUSIPCO-86.

[2] G M KUHN and K OJAMAA, 'Automatic recognition of words differing in distinctive quantity', Proceedings of the 11th International Congress of Phonetics Sciences, Estonia, USSR, August 1987.

[3] S E LEVINSON, 'Continuously variable duration hidden Markov models for automatic speech recognition', Computer Speech and Language, 1, pp29-46, (1986).

[4] L R RABINER and S E LEVINSON, 'A speaker-independent, syntax driven, connected word recognition system based on hidden Markov models and level building', IEEE Trans. Acoustics, Speech and Signal Processing, ASSP-3, 3, pp561-573, (1985).

[5] M J RUSSELL and A E COOK, 'Experimental evaluation of duration modelling techniques for automatic speech recognition', Proceedings of ICASSP-87, Dallas, Texas, pp2376-2379, (1987).

[6] M J RUSSELL, R K MOORE, M J TOMLINSON and J C A DEACON, 'RSRE speech database recordings 1983: part II. Recordings made for automatic speech recognition assessment and research', RSRE report no. 84008, (1984).

[7] M J RUSSELL and R K MOORE, 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', Proceedings of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-85, Tampa, Florida, pp5-8, (1985).

[8] M J RUSSELL, 'Maximum likelihood hidden semi-Markov model parameter estimation for automatic speech recognition', RSRE memorandum number 3837, (1985).

[9] J S BRIDLE and N C SEDGWICK, 'A method for segmenting acoustic patterns with applications to automatic speech recognition', Proceedings of IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP-77, pp656-659, (1977)

[10] L A LIPORACE, 'Maximum likelihood estimation for multivariate observations of Markov sources', IEEE Trans. Information Theory, IT-28, pp729-734, (1982).

[11] L R RABINER and B H JUANG, 'An introduction to hidden Markov models', IEEE ASSP Magazine, pp4-16, January 1986

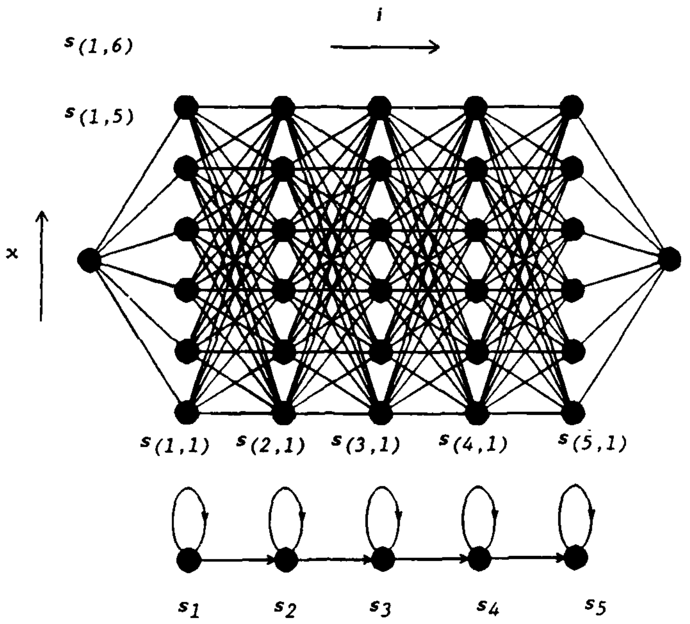


FIGURE 1: Diagram showing the relationship between a simple left−right 5 state HMM and its associated CDHMM. In this case the maximum state duration D_{max} is 6.

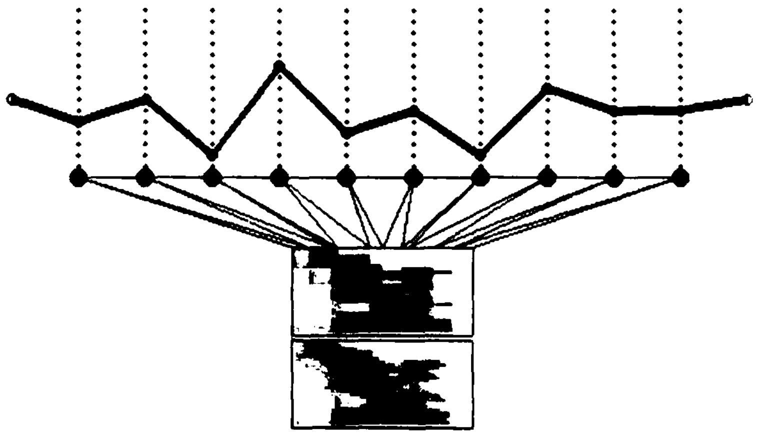


FIGURE 2: Diagram showing the relationship between a HMM state sequence and a path through the associated CDHMM. The spectrogram at the bottom of the figure represents an example of the digit 'zero'. The top spectrogram is a synthesised version of 'zero' which represents the best explanation of the bottom spectrogram in terms of the HMM. The cones show how the states of the HMM map onto the two patterns.

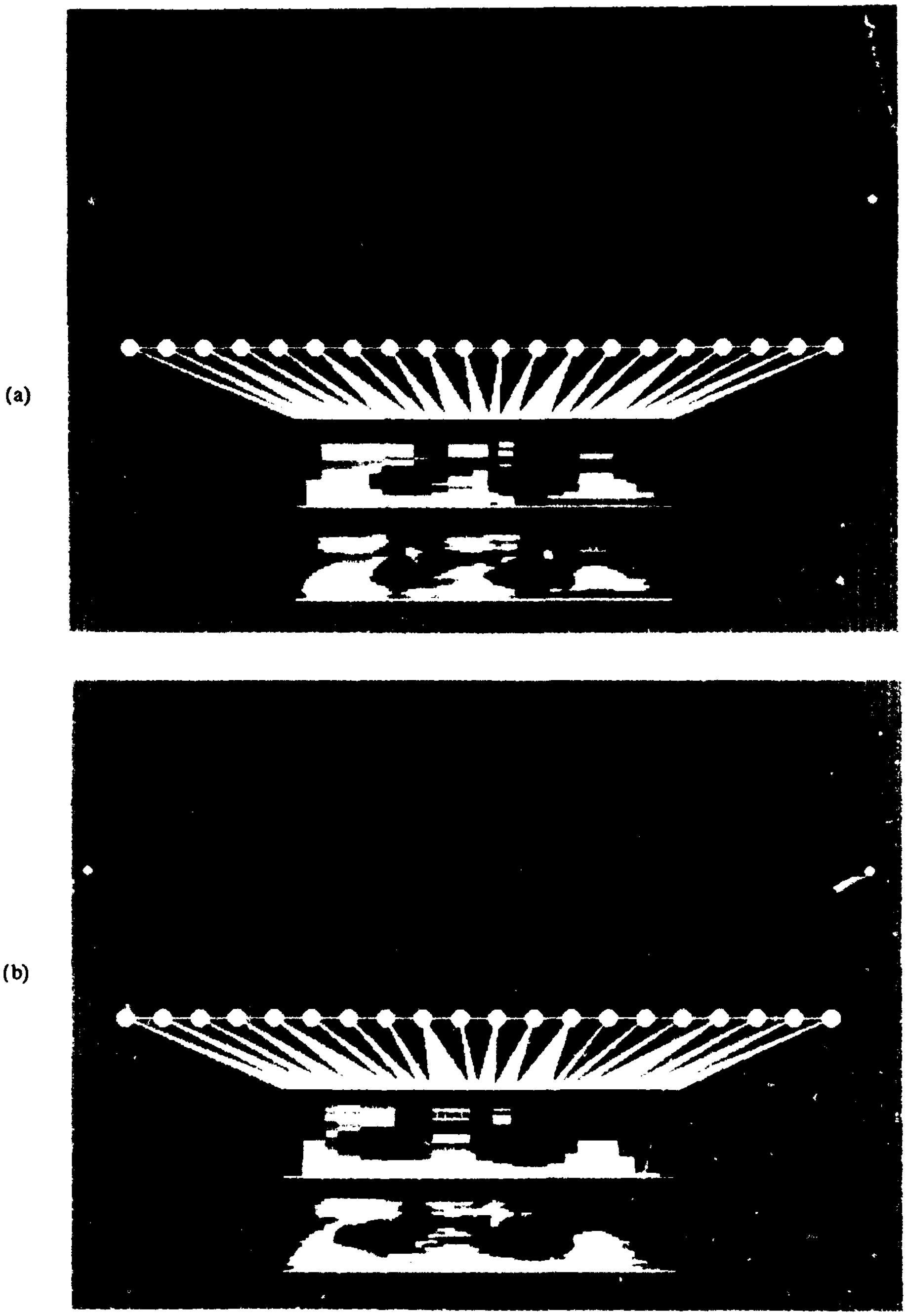
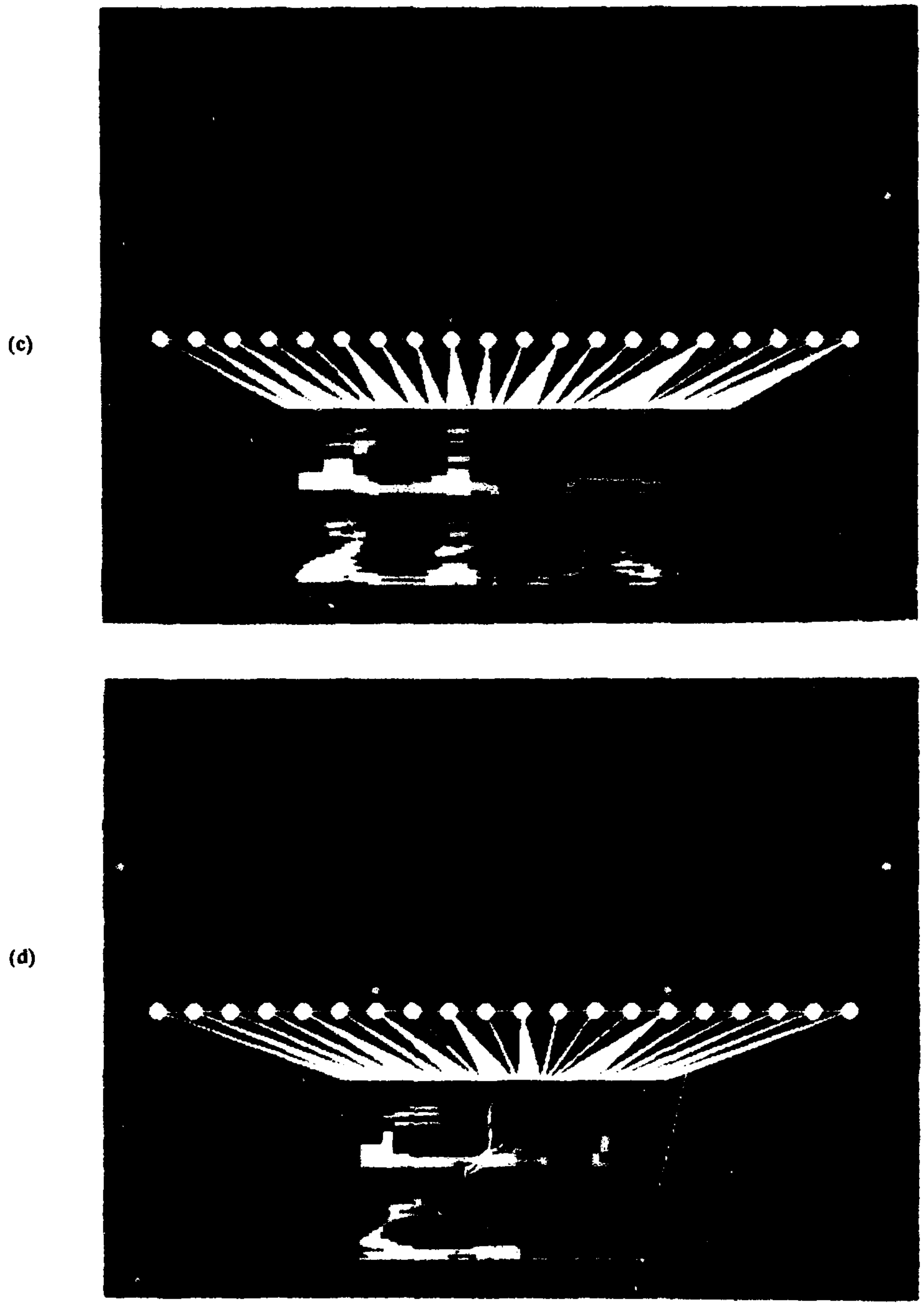


FIGURE 3: Representations of the 20 state CDHMMs of the sentence 'Why are you early, you owl?' used in the experiments. Speakers RM (a), SJ (b), HP (c) and CR (d). For each speaker the bottom spectrogram represents an example of the sentence spoken at normal rate, while the top spectrogram is the best explanation of the botom spectrogram in terms of the HMM. The cones show how the states map onto the two patterns.



(FIGURE 3: Continued).

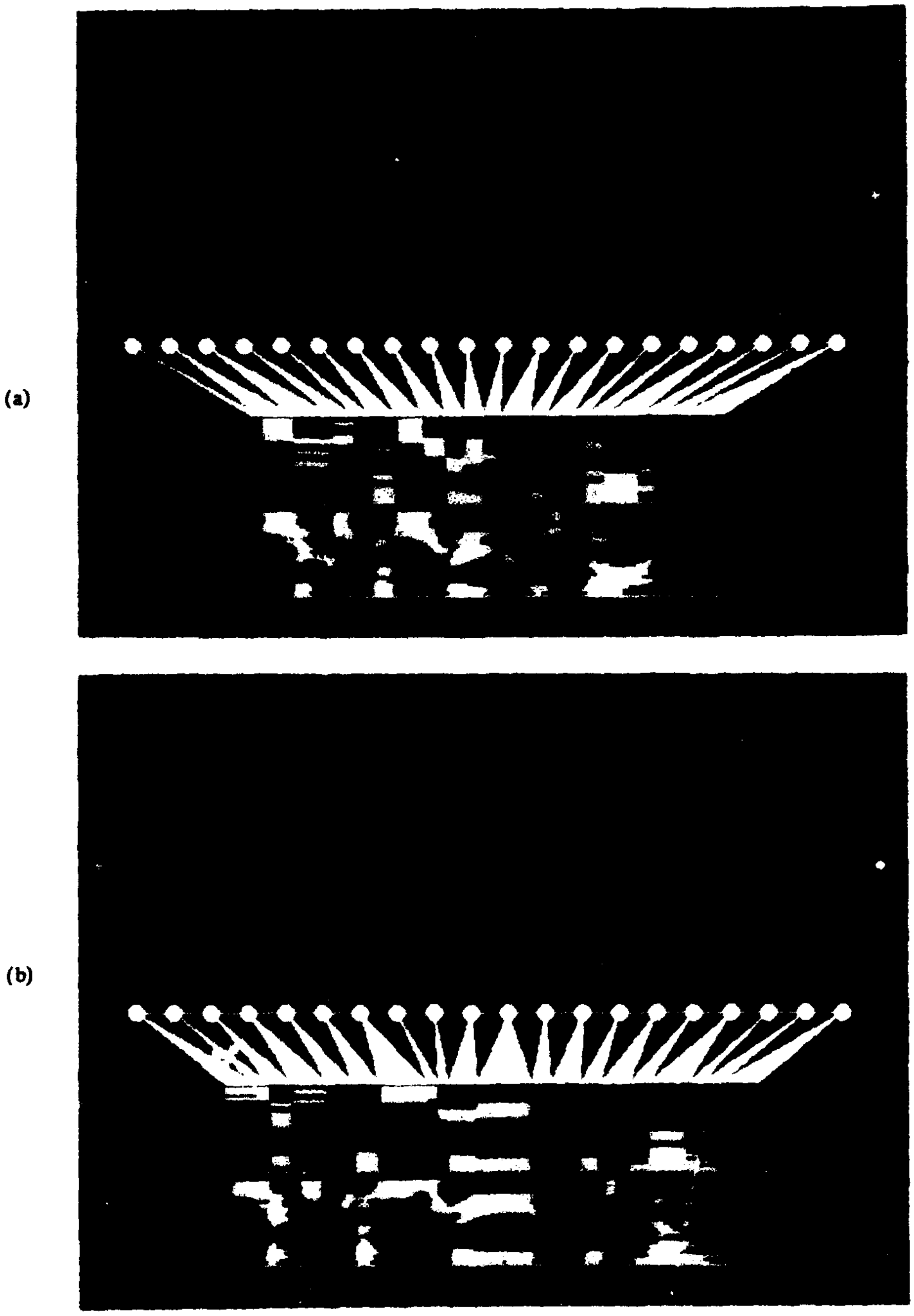
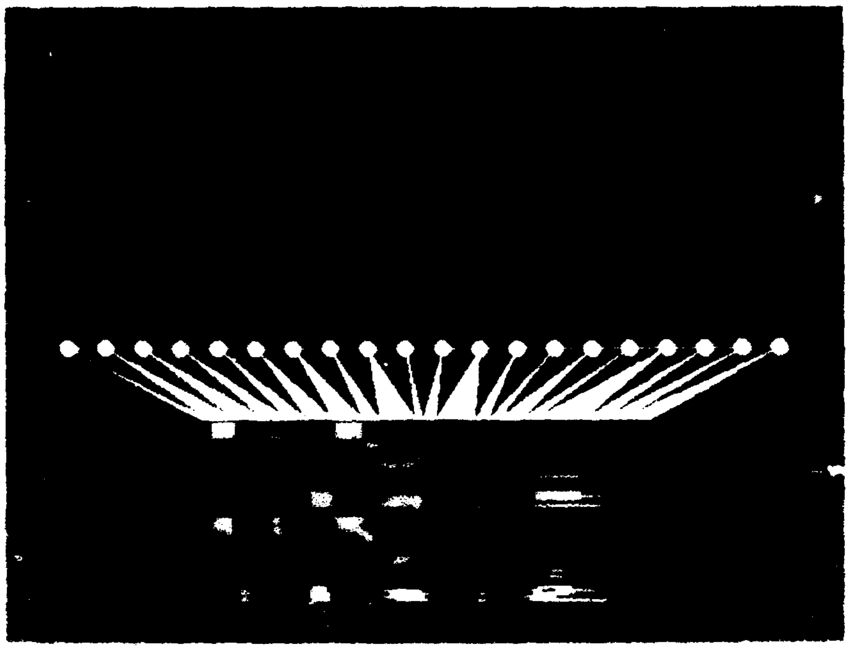
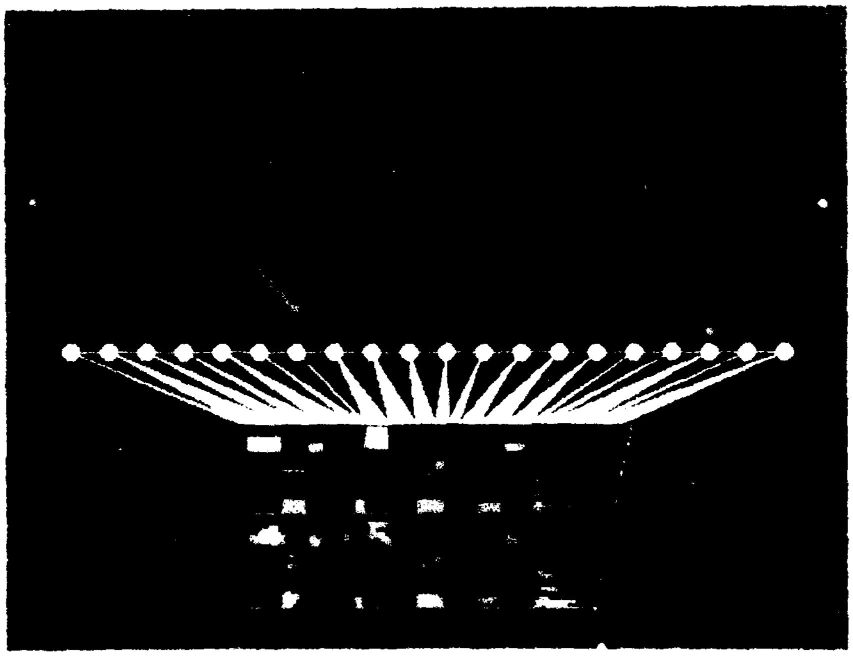


FIGURE 4: Representations of the 20 state CDHMMs of the sentence 'Six plus three equals nine' used in the experiments. Speakers RM (a), SJ (b), HP (c) and CR (*d*). For each speaker the bottom spectrogram represents an example of the sentence spoken at normal rate, while the top spectrogram is the best explanation *of the botom spectrogram in terms of the HMM*. The cones show how the states map onto the two patterns.

(c)



(d)



(FIGURE 4: Continued).

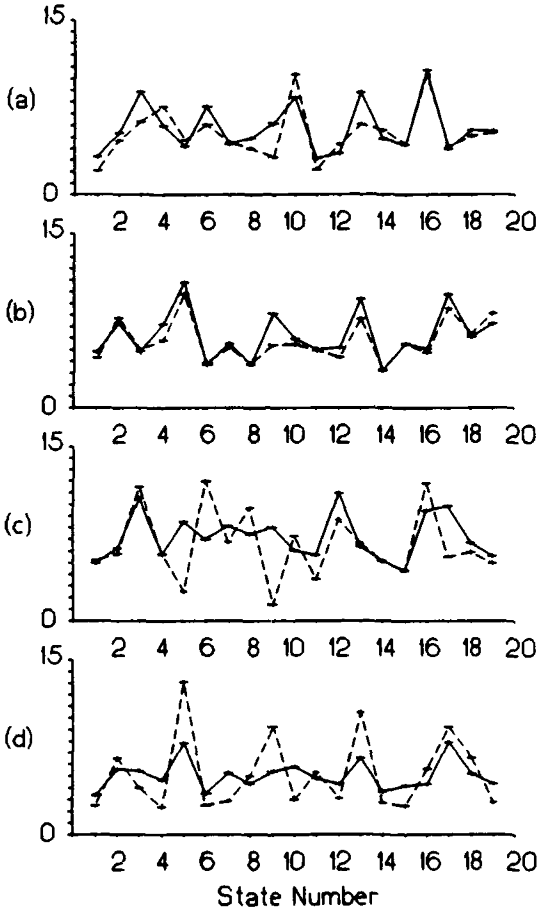


FIGURE 5: Expected state duration as a function of state number for normal (———) and fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Why are you early, you owl?'.

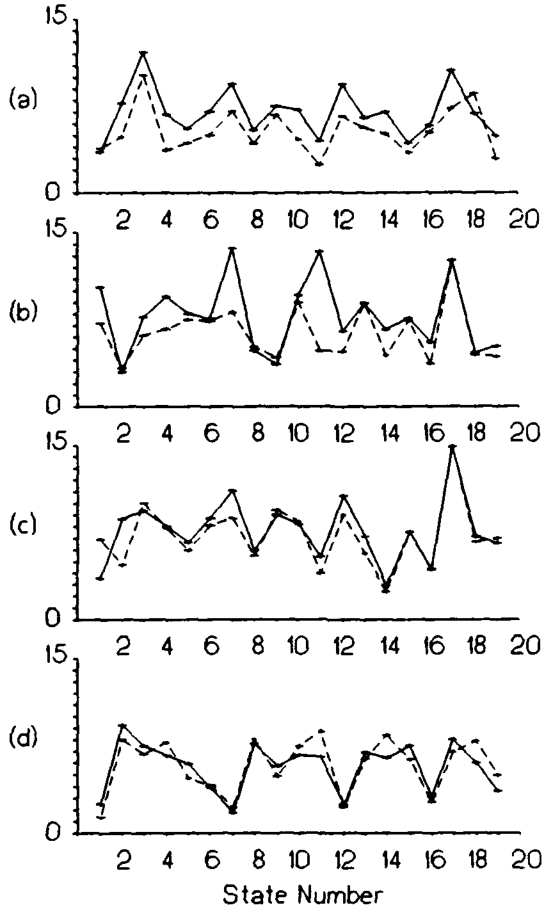


FIGURE 6: Expected state duration as a function of state number for normal (———) and fast (– – – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d; and the sentence 'Six plus three equals nine'.

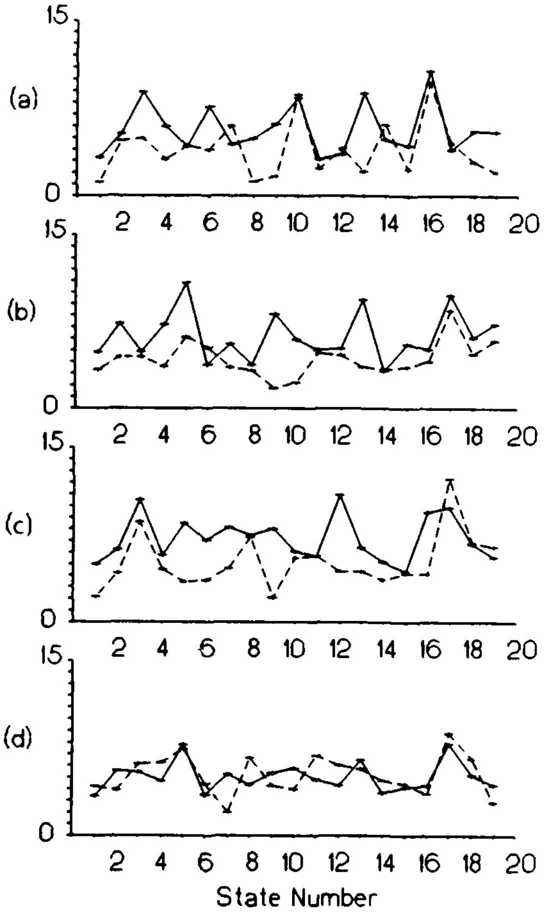


FIGURE 7: Expected state duration as a function of state number for normal (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Why are you early, you owl?'.

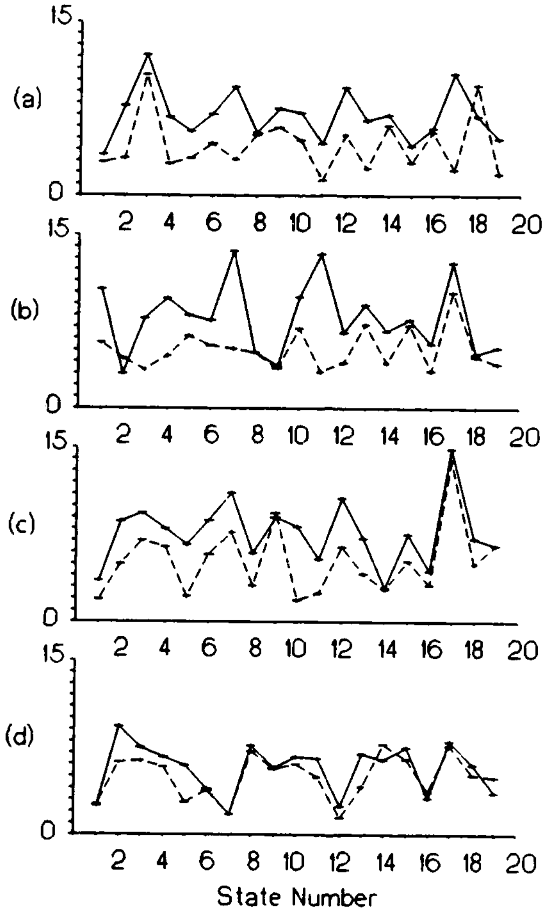


FIGURE 8: Expected state duration as a function of state number for normal (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Six plus three equals nine'.

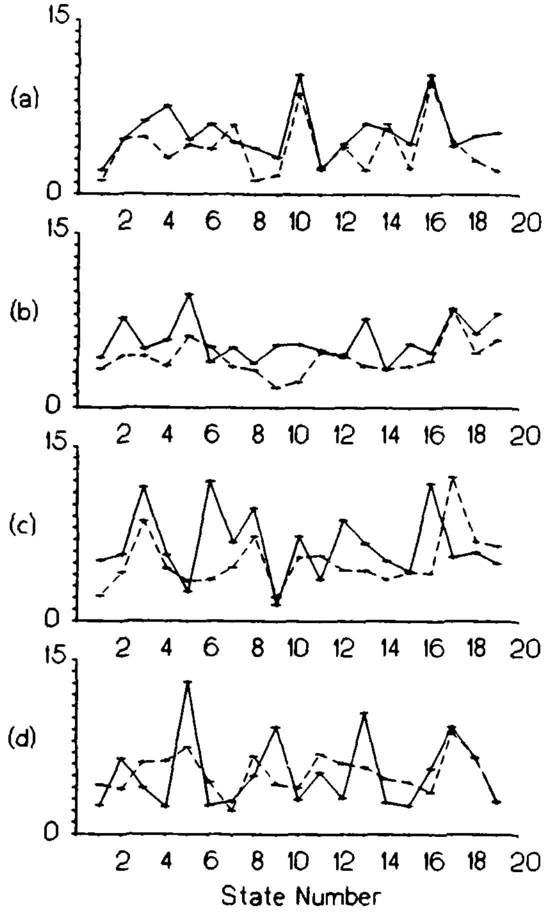


FIGURE 9: Expected state duration as a function of state number for fast (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Why are you early, you owl?'.

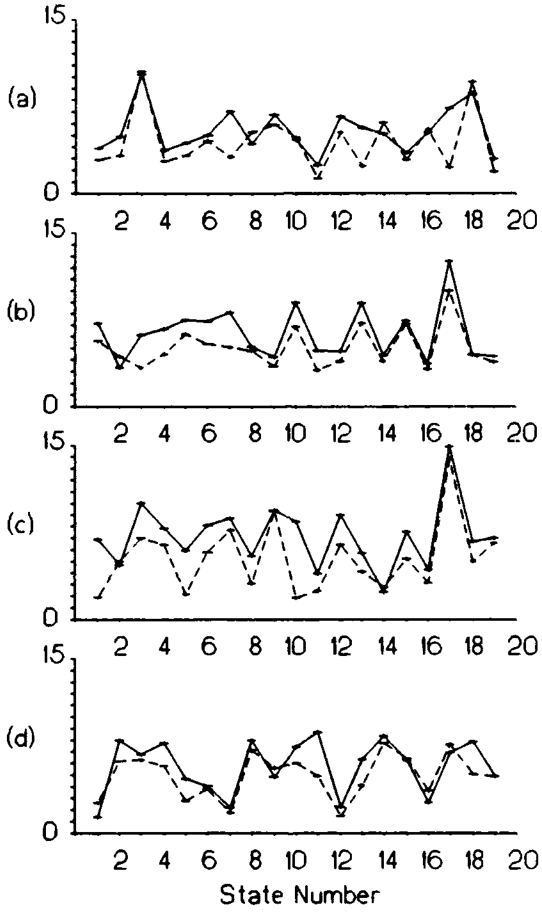


FIGURE 10: Expected state duration as a function of state number for fast (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) ard CR (d) and the sentence 'Six plus three equals nine'.

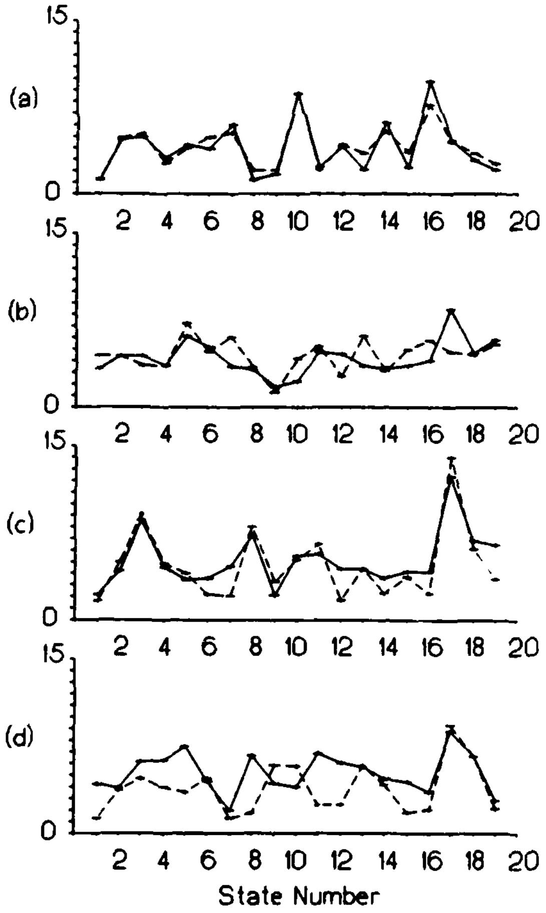


FIGURE 11: Expected state duration as a function of state number for very fast (———) and fast with meter (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Why are you early, you owl?'.

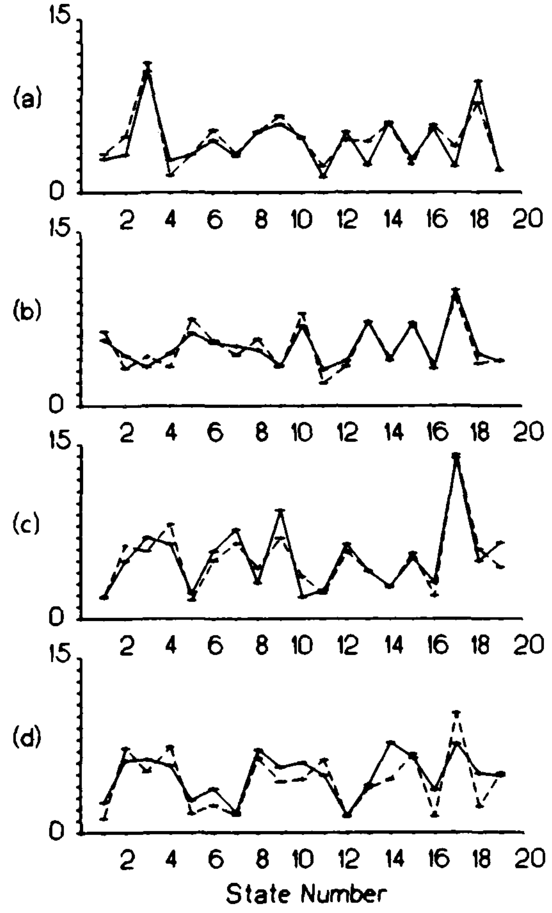


FIGURE 12: Expected state duration as a function of state number for very fast (———) and fast with meter (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Six plus three equals nine'.

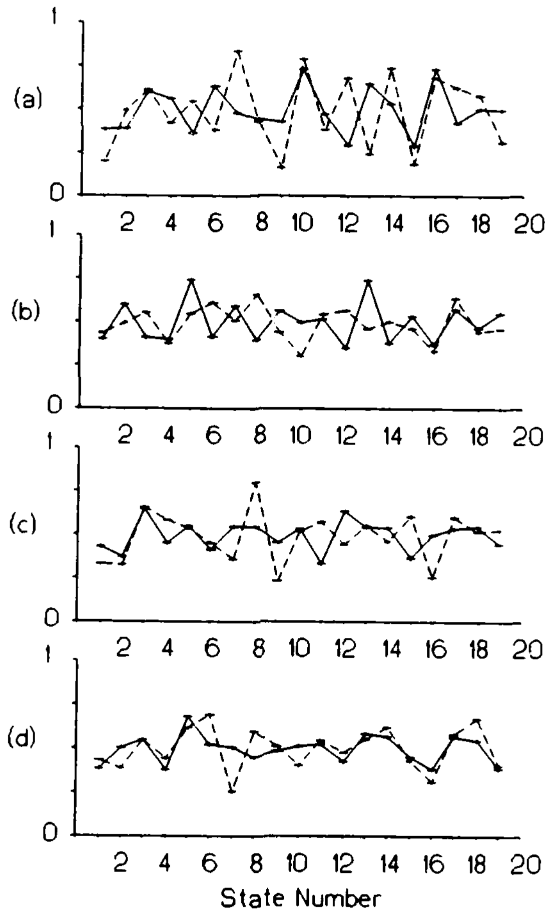


FIGURE 13: Expected state duration ratio as a function of state number for normal (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Wny are you early, you owl?'.

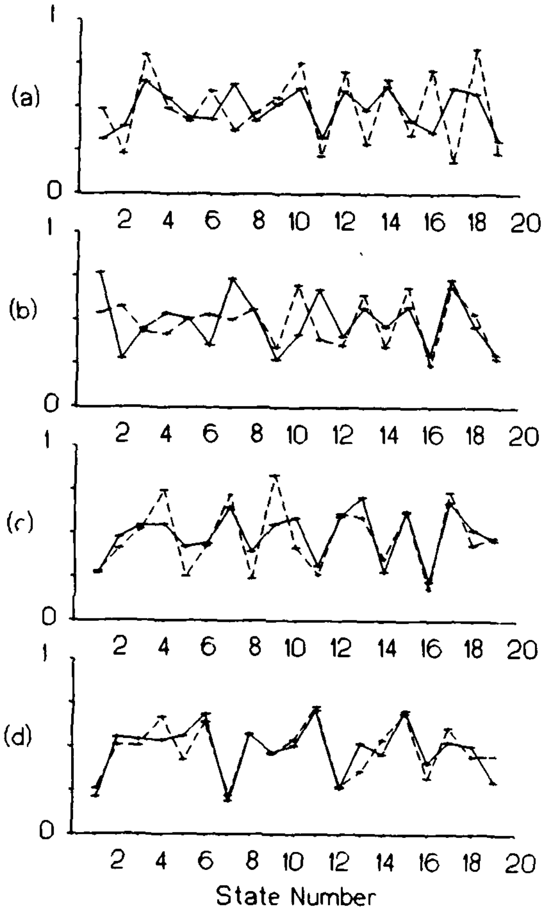


FIGURE 14: Expected state duration ratio as a *function* of state number for normal (———) and very fast (– – – – –) rate speech. The graphs are for 20 state CDHMMs, speakers RM (a), SJ (b), HP (c) and CR (d) and the sentence 'Six plus three equals nine'.

APPENDIX 1

The SPAR list of sentences used in the experiments

1. George made the girl measure a good blue vase

2. Be sure to fetch a file and send their's off to Hove

3. Six plus three equals nine

4. Kathy hears a voice amongst SPAR's data

5. Why are you early, you owl?

DOCUMENT CONTROL SHEET

Overall security classification of sheet UNCLASSIFIED

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference Memorandum 4152	3. Agency Reference	4. Report Security Classification Unclassified

5. Originator's Code (if known) 7784000	6. Originator (Corporate Author) Name and Location Royal Signals and Radar Establishment St Andrews Road, Malvern, Worcestershire WR14 3PS
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location

7. Title

EXPLICIT MODELLING OF STATE DURATION CORRELATIONS IN HIDDEN MARKOV MODELS

7a. Title in Foreign Language (in the case of translations)

7b. Presented at (for conference papers) Title, place and date of conference

8. Author 1 Surname, initials Russell M J	9(a) Author 2 Sime L	9(b) Authors 3,4...	10. Date 9.88	pp. ref. 26

11. Contract Number	12. Period	13. Project	14. Other Reference

15. Distribution statement

Unlimited

Descriptors (or keywords)

continue on separate piece of paper

Abstract

See Overleaf:

ABSTRACT

In recent years considerable effort has been directed towards improving the treatment of durational structure in hidden Markov model (HMM) based approaches to speech pattern modelling. In general these studies have been concerned with more accurate modelling of the variations in segment duration which occur when words are spoken at a nominally constant speaking rate. However, recent work has shown that some of the performance gains which can be achieved by improved duration modelling are lost when the words in the test set are spoken at a different rate from those in the training set.

This memorandum presents an approach to solving this problem based on the capture and use of information about state duration correlations. A method for measuring correlations between the durations of adjacent states in a HMM is described. The method involves expanding a standard HMM into a special type of hidden semi−Markov model (HSMM), called a Correlated Duration HMM (CDHMM), in which each state of the original HMM is expanded into a set of fixed−duration HSMM states. The probabilities associated with transitions between these states are measures of state duration correlation.

Experiments are described in which the CDHMM method is applied to a set of sentences spoken at four different speaking rates.